International Conference on Education in Mathematics, Science and Technology

www.icemst.com March 24-27, 2022 Antalya, TURKEY www.istes.org

Comparison of Two Exam Evaluation Methods for Objectivity

Biserka Kolarec

The University of Zagreb, Faculty of Agriculture, Croatia, https://orcid.org/0000-0003-1434-3533

Marina Ninčević

The University of Zagreb, Faculty of Agriculture, Croatia, https://orcid.org/ 0000-0003-2316-2072

Abstract: The object of research is a statistics exam that contains problem tasks. One examiner performed two exam evaluation methods to repeatedly evaluate the exam. The goal was to compare the methods for objectivity. One of the two exam evaluation methods we call a serial evaluation method. The serial evaluation method assumes evaluation of all exam tasks of an individual student in sequential order, so evaluation "student by student". Unlike that, a parallel evaluation method assumes the "task by task" evaluation of exams for the whole group of students. A paired samples analysis of exam results indicates a statistically significant difference between methods. Further analysis showed a statistically significant difference in results obtained by the serial evaluation and the repeated serial evaluation, while the difference in results obtained by the parallel evaluation and the repeated parallel evaluation turned out not to be statistically significant. Furthermore, our research gave evidence that the repeated parallel evaluation changes result significantly less than the repeated serial evaluation. Consequently, we consider the parallel evaluation a more objective tool in exam evaluation than the serial evaluation.

Keywords: Exam evaluation, Serial evaluation, Parallel evaluation, Paired samples analysis

Introduction

Nowadays, teachers employ active learning activities such as flipped classrooms, peer learning, group work activities, collaborative argumentation, formative assessment (see Latifi & Noroozi, 2021; Latifi et al., 2020, 2021; Noroozi 2018, 2022; Noroozi et al., 2016; 2020; Valero Haro et al., 2019; 2022). This is also the case with written exams. While Ortega-Sanchez (2016) poses a question on how effectively are we using written exams, here we confront two methods in the critical analysis of how objectively do we evaluate written exams. Certainly, each teacher spends a lot of time designing a written exam and balancing different levels of difficulty of the exam tasks to distinguish between low and high levels of students' knowledge. For all that not to be in vain, attention must also be paid to methods that lead to better objectivity in the exam evaluation. When we searched for literature on this subject, we hardly found some. Lack of literature on the topic indicates we pay

Kolarec, B., & Ninčević, M. (2022). Comparison of Two Exam Evaluation Methods for Objectivity. In P. Dankers, M. Koc, & M.L. Ciddi (Eds.), *Proceedings of ICEMST 2022-- International Conference on Education in Mathematics, Science and Technology* (pp. 158-168), Antalya, TURKEY. ISTES Organization.

very little attention to the subject of objectivity in exam evaluation. How come this question on subjectivity/objectivity of exam evaluation is not a subject of much research? In this research, we hope to throw a little light to the dark corner of the subject of objectivity in exam evaluation.

Most opinions on exam evaluation emphasize its formative role: to communicate and inform teaching and improve learning (Liljedahl, 2010, Mitchell and Neill, 1992). But, it is rare in the practice of exam evaluation in higher education. When a university teacher constructs an exam, it is with a prime goal to measure student achievements and assign them grades. So, in the practice of higher education, exam evaluation plays mostly a normative role. Therefore, exam evaluation must be objective (McTighe, Ferrara, 1998).

According to Baehr (2004) there are four steps in the process of exam construction and evaluation: 1) to define learning outcomes that will be tested, 2) to have in mind evidence of learning outcomes achievement, 3) to set up a scale of points such that all learning outcomes are adequately scored, and 4) to assign grades through objective exam evaluation. We focus on the last step. Although in 1975 L. J. Herbst made the statement: "*70s are likely to be a decade in which objective testing will be firmly established in higher education,"* today we still struggle with objective exam evaluation. For example, Romagnano (2001) states: "*Objectivity, like the mythical pot of gold at the end of the rainbow, would be wonderful if we could have it, but it does not exist."* Further, Liljedahl (2010) claims that all assessments of students' mathematical understanding are subjective. We agree with both and nevertheless search for an objective exam evaluation method.

Literature Review

The matter of objectivity in exam evaluation is rarely addressed in scientific research. Mostly, authors state the necessity of objective exam evaluation (E. Hoosain and B. Naraine, 1999), but there is a lack of instructions on concrete actions to achieve objectivity. One research that addresses the subject was done in 2011 by Kriauziene, Krylovas, and Kosareva. In the research student's exams were independently verified by six teachers and some disagreement in results was found. Further, L. J. Herbst (1975) reports that even "*repeated marking of the paper* ***by the same examiner*** *is likely to show significant variability*," referring to essay-type exams. Contrary to them, math exams are considered far more objective. Indeed, the evaluation of deterministic tasks like "*2+2=*" doesn't depend on any subjective judgment of a teacher, but things change in exams that contain problem tasks. White (2019) distinguish between objective and subjective exams; objective exams have either a right or a wrong answer, while subjective have answers in-between the right and the wrong answer. Exams that contain problem tasks are definitely subjective. Further, the evaluation of mathematical exams with problem tasks tends to be subjective.

To evaluate problem tasks, one must have clear scoring lists organized in rubrics. That is exactly what Guat Poh et al. (2015) emphasize when they suggest creating a marking scheme rubric. Evaluators should indeed have details of scoring of characteristics steps of a solution to credit, not just exact solution, but also student's attempt

to solve the problem. But, if there are many ways to the correct solution, even rubrics can become incomplete, and then scoring must be adapted *ad hoc* to an individual case one faces. Firm, detailed and comprehensive scoring criteria are, indeed, necessary, but not at all sufficient to ensure objectivity in exam evaluation. We claim that the choice of the exam evaluation method also plays a substantial role in the objectivity of evaluation.

Evaluation Methods Description

We use to grade exams "student by student". It means to take an individual student exam sheet, go through it in its entirety, and assign points to each task according to prescribed criteria before passing to another exam sheet. We name this procedure a *serial evaluation.* It assumes going through the entire evaluation criteria over and over again.

From time to time we noticed that we scored the same solutions with a different number of points. So, we spotted subjectivity in exam evaluation. Then we started to wonder: maybe the evaluation of just one task at a time for the whole group would lead to better consistency in results, i.e. to higher objectivity? Namely, it is a common practice in team evaluation that a single examiner grades a single problem (or a few) for all students rather than the whole exam for a subgroup of students. The reason for this is obvious: to reduce the impact of different criteria or grading practices examiners have. The question is: can we trust more the objectivity of just one examiner?

The principle of the evaluation method that we call *a parallel evaluation* is exactly the evaluation of exams "task by task". In the parallel evaluation, just one task at a time is evaluated for the whole group. Practically, it involves grading each student's submission for a single task before moving to the next.

We are confident that teachers use both, serial and parallel evaluation methods extensively. However, the search for the terms "the serial evaluation" and "the parallel evaluation" gave results only in the fields of medicine and data mining. We choose names for evaluation methods miming terms in electrical circuits (connecting resistors, for example), based on the order of steps performed in the evaluation procedure (illustrated in Figure 1).

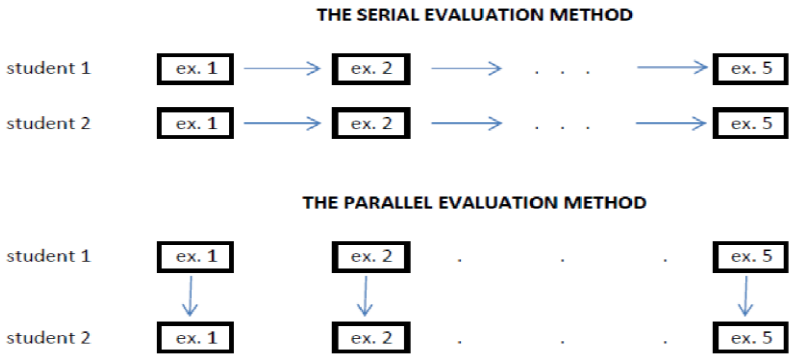

Figure 1. Schemes of Steps performed in the Serial and the Parallel Evaluation

The serial evaluation assumes looking at scoring criteria of all tasks for each exam sheet. If one meets some unusual solution, a new decision on scoring must be made. In the parallel evaluation, a teacher has just one evaluation criterion in mind at a time, and the scoring criterion stays firm for the whole group. If the examiner deals with just one task at a time, he/she can spot more easily different ways to the correct solution and then adopt scoring criteria to include them all.

Research Subject, Methodology, and Hypotheses

The subject of the research was one exam in a university statistics course. It consisted of five problem tasks on descriptive statistics and the basics of probability theory. There were a total of 35 students that year, and we included all their exams in the research. Exam sheets of all students were evaluated by the same evaluator four times over nearly one and a half years, two times with serial and two times with parallel evaluation method. The time gap between evaluations was long enough to ensure that the evaluator recalls no previous scoring.

The research was performed as follows. The initial idea was to see if there is a statistically significant difference in the results of the serial and the parallel evaluation methods. The exam in question was already evaluated with the serial evaluation method and the results were available. By results, we mean the total number of points for an individual student ranging from 0 to the maximum of 40 points. Approximately half a year later the evaluator evaluated the exam with the parallel evaluation method. Old scoring criteria were available and reused, and before that evaluation, the third party made all previous scoring invisible. Obtained data gave evidence that there is a statistically significant difference among results.

Next, we wanted to check how the results of both methods change over time. For that purpose, the examiner repeated evaluation with both methods. The repeated serial evaluation was performed four months after the parallel evaluation and the repeated parallel evaluation some five months after the repeated serial evaluation. We ensured a sufficient time gap to prevent any possibility of recollection of the previous scoring. Also, before each evaluation, the same precaution measures were applied as before to guarantee that no previous results are visible. The same scoring criteria were used repeatedly in all evaluations.

Scoring the Problem Task – an İllustration

To check students understanding of the classical probability definition, the following task was given:

> "*Dice is thrown two times. From the obtained numbers one forms a fraction: the number that fell first becomes a numerator and the number that fell second is a denominator. What is the probability of obtaining a reducible fraction?"*

The correct answer is 13/36 because among 36 possible fractions there are 13 reducible ones. To obtain the probability, students were expected to list and count reducible fractions and divide the obtained number with 36, the number of all possible fractions. As one could expect, some lists were incomplete. According to the scoring

criteria, the correct solution was assigned with eight points, and this maximum was lowered by one for each missing or extra fraction; however, the maximal reduction was four points. Further, for very poor lists students could get two points if they wrote the number of all possible fractions and one more if they wrote probability as a fraction. Below are some pictures to illustrate different solutions: there is the correct solution and partially correct ones with different degrees of accuracy given in Figure 2.

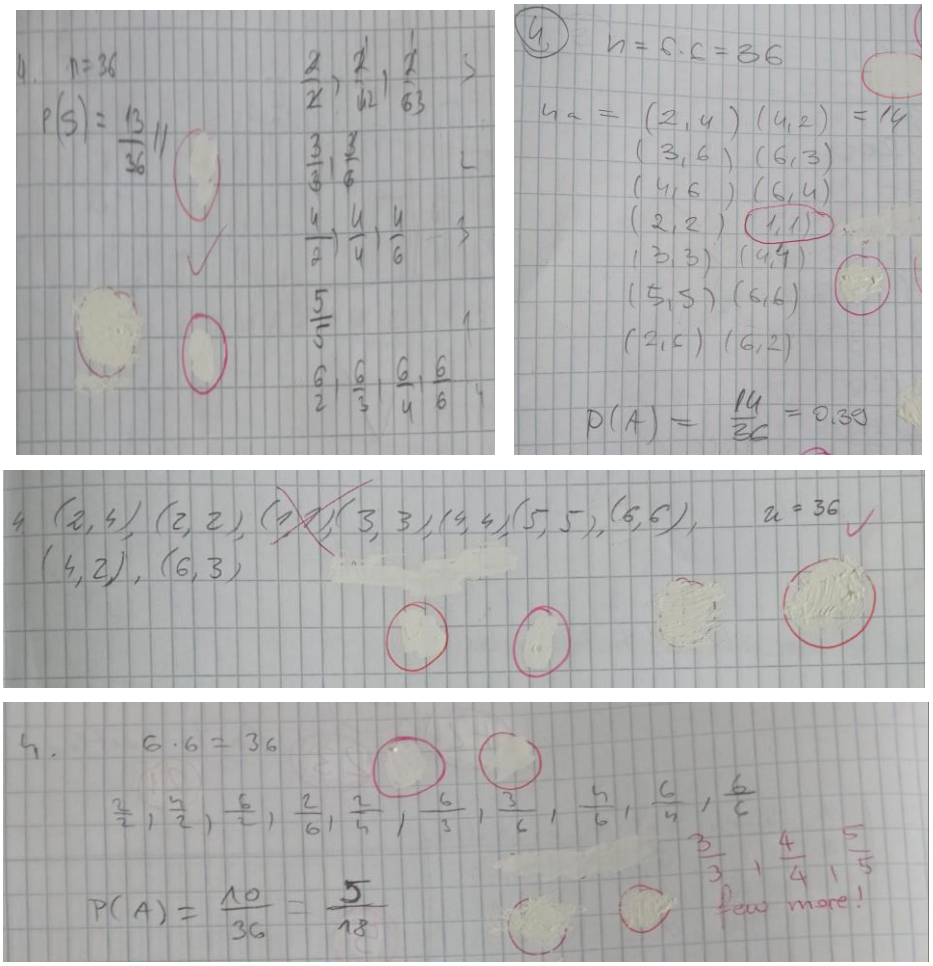

Figure 2. Some Pictures of Problem Solutions (Empty circles on each picture witness removal of old scoring in each new evaluation.)

Data set

At the end of all evaluations, we had four sets of exam evaluation results obtained by the use of the serial evaluation (S), the parallel evaluation (P), the repeated serial evaluation (RS), and the repeated parallel evaluation (RP). Having all those, we further calculated absolute values of differences between results of the serial and the repeated serial evaluation (|S-RS|), and absolute values of differences between results of the parallel and the repeated parallel evaluation (|P-RP|). Namely, the order of first and second evaluation of the same type is of no importance, one is interested only in absolute differences. The data set is given in Table 1.

Table 1. Student Results obtained by the Serial (S), the Repeated Serial (RS), the Parallel (P), the Repeated Parallel Evaluations (RP) with Absolute Values of Paired Differences (|S-RS|) and (|P-RP|).

Student	S	P	RS	RP	|S-RS|	|P-RP|
1	35	30	32	31	3	1
2	32	30	30	29	2	1
3	30	27	27	27	3	0
4	24	21	23	23	1	2
5	36	33	33	33	3	0
6	30	30	29	29	1	1
7	41	38	39	39	2	1
8	24	23	23	24	1	1
9	32	28	26	26	6	2
10	23	18	18	18	5	0
11	25	19	23	20	2	1
12	26	22	22	25	4	3
13	39	37	38	37	1	0
14	23	23	26	24	3	1
15	34	27	31	31	3	4
16	22	22	21	21	1	1
17	32	32	31	31	1	1
18	33	25	25	26	8	1
19	31	29	32	32	1	3
20	31	32	31	31	0	1
21	32	25	25	26	7	1
22	31	29	31	31	0	2
23	34	34	33	33	1	1
24	17	15	19	16	2	1
25	31	31	30	32	1	1
26	19	18	18	19	1	1
27	38	33	32	29	6	4
28	25	21	21	25	4	4
29	29	27	30	29	1	2
30	32	30	34	30	2	0
31	10	11	11	10	1	1
32	34	34	34	32	0	2
33	34	29	31	31	3	2
34	34	34	34	34	0	0
35	16	16	17	16	1	0
mean	29.11	26.66	27.43	27.14	2.31	1.34

The data set is a typical example of paired samples since each data refers to the same individual so we chose paired sample analysis to check the following four hypotheses:

H_1 – there is no statistically significant difference between results obtained by the serial and the parallel evaluation.

H_2 – there is no statistically significant difference between results obtained by the serial evaluation and the repeated serial evaluation.

H_3 – there is no statistically significant difference between results obtained by the parallel evaluation and the repeated parallel evaluation.

H_4 – absolute differences between results obtained by the parallel evaluation and the repeated parallel evaluation are significantly greater than the absolute differences between results obtained by the serial evaluation and the repeated serial evaluation.

The last hypothesis was set up to see if the repeated parallel evaluation changes result significantly less over time compared to the parallel evaluation than the repeated serial evaluation compared to the serial evaluation.

Results

Tests of all hypotheses were performed in "The R Project for Statistical Computing" at the 5% significance level. For each test all pairs: S and P, S and RS, P and RP, as well as |S-RS| and |P-RP|, form typical paired sample data. Therefore, we performed a paired t-test for the equality of means to test all hypotheses. To check the paired t-test assumption that the differences of the matched pairs follow a normal probability distribution, we used appropriate normal Q-Q plots (Figure 3).

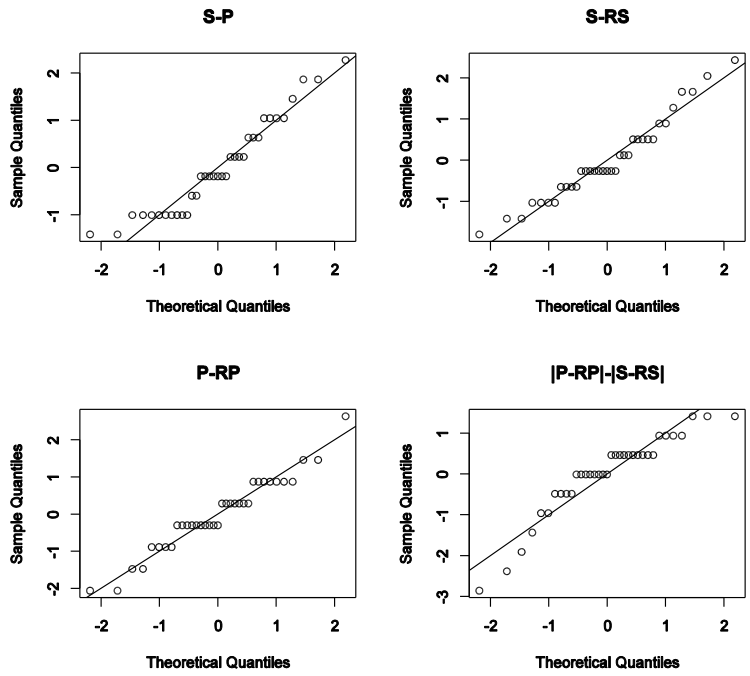

Figure 3. Normal Q-Q Plots for Differences of Results in All Four Tests

The 5% significance level in paired sample tests gives the two-tail critical point of 2.03. For H_1, the test showed that the null hypothesis should be rejected because the test statistic was 5.9538, with a p-value of 9.888×10^{-7}. There is a strong indication of a statistically significant difference in the results obtained by the serial and the parallel evaluation.

To test hypothesis H_2, we performed a paired t-test for the equality of means for results obtained by the serial evaluation and the repeated serial evaluation. The test statistics 3.8382 was greater than the two-tailed critical point 2.03 with the p-value of 0.00051. So, the difference in the results of the serial evaluation and the repeated serial evaluation is statistically significant.

A paired t-test for the equality of means for the results obtained by the parallel evaluation and the repeated parallel evaluation showed that data are in favor of the null hypothesis, i.e. there is no statistically significant difference in the results of the parallel evaluation and the repeated parallel evaluation. Namely, the test statistics -1.6862 belong to the acceptance area [-2.03, 2.03], with a p-value of 0.1009.

To test hypothesis H_4, we performed a paired t-test on the null hypothesis that the mean of the absolute differences of the parallel evaluation and the repeated parallel evaluation is greater than the mean of absolute differences of the serial evaluation and the repeated serial evaluation. The test suggested that this null hypothesis should be rejected because the test statistic -2.7273 is smaller than the one tail critical point -1.69 with the p-value of 0.005. So, we conclude that the absolute differences of the results of the parallel and the repeated parallel evaluation are significantly less than the absolute differences of the results of the serial evaluation and the repeated serial evaluation.

For easier visibility, a row with the means of all samples is provided in the last row of Table 1 above. Relations between means are visible the best on the boxplots of samples given in Figure 4.

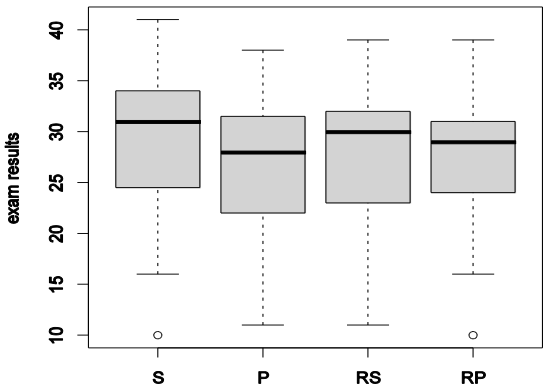

Figure 4. The Comparison of Boxplots of S, P, RS, and RP Data Sets

Figure 5 gives the boxplots of absolute differences |P-RP| and |S-RS|. One can observe smaller variability of absolute differences of results of the parallel and the repeated parallel evaluation methods.

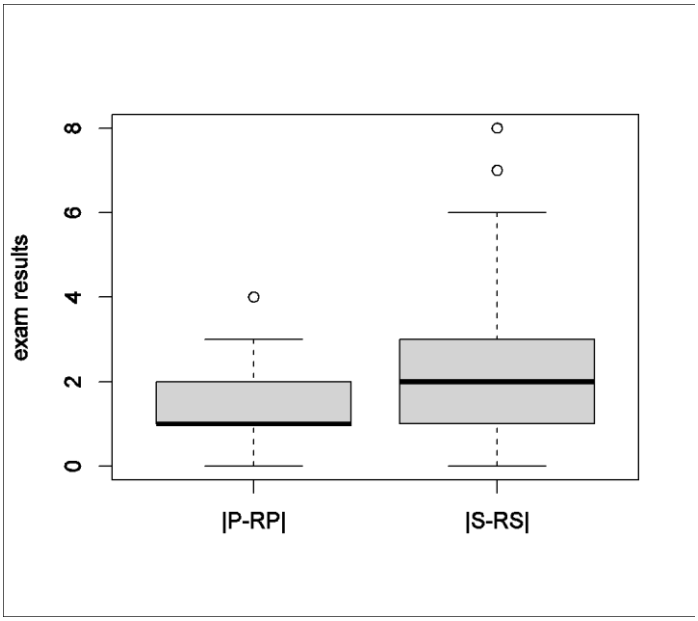

Figure 5. The Comparison of Boxplots of |P-RP| and |S-RS| Data Sets

Discussion and Conclusion

In higher education, there are high stakes in exam results. Therefore, every teacher should aim at consistent and objective exam evaluation. The objectiveness of exam evaluation is not at all an issue in exams that contain only tasks of an objective type, but it becomes a great issue in exams that contain problem tasks.

This research was done to compare two exam evaluation methods, the serial and the parallel evaluation method for objectivity exactly on the exam with problem tasks. Data give evidence of a statistically significant difference in exam results obtained by the two methods, so we consider them different. Further, data show no statistically significant difference in the results of the parallel evaluation and the repeated parallel evaluation, but they do establish the existence of a statistically significant difference in the results of the serial evaluation and the repeated serial evaluation. Further, statistical analysis confirms that the absolute differences of the results of the parallel and the repeated parallel evaluation are significantly less than the absolute differences of the results of the serial evaluation and the repeated serial evaluation. Consequently, we conclude that the parallel evaluation leads to more objective results compared to the serial evaluation.

We argue that, although judgments and decisions one makes in the exam evaluation are inevitably subjective, objectivity can be increased by the use of the parallel evaluation method. In our experience, the parallel evaluation takes more time than the serial evaluation, but it is a small price to pay to achieve higher objectivity. Anywhere applicable, we recommend the evaluation of exams with the parallel evaluation method.

We are aware of the fact that this research alone just starts a discussion on objectiveness in exam evaluation. Results obtained here are yet to be confirmed in more thorough research. Although our research was done on a mathematical exam, we are confident that conclusions stay valid for any exam with problem tasks. Hopefully, this research will start a discussion on ways to ensure higher objectivity in exam evaluation, a topic far, far under-discussed in literature and scientific research.

References

Baehr M. (2004): Evaluation Methodology, *Pacific Crest* (630), 737-1067

Guat Poh B. L., Muthoosamy K., Lai C. C., Hoe G. B. (2015). A Marking Scheme Rubric: To Assess Students' Mathematical Knowledge for Applied AlgebraTest, *Asian Social Science* 11 (24), 18-24

Herbst, L. J. (1975): Objective testing in higher education, *The Vocational aspect of Evaluation* 27 (66), 21-25

Hoosain, E. & Narraine, B. (1999): Evaluation in the mathematics classroom, *Humanistic Mathematics Network Journal* 20, 11-13

Kriauziene, R., Krylovas, A. & Kosareva, N. (2011): Subjectivity Problem in Student Assessment: Theoretical and Practical Aspects, *Social Technologies* 1(1), 121-138

Latifi, S., & Noroozi, O. (2021). Supporting argumentative essay writing through an online supported peer-review script. Innovations in Education and Teaching International, 58(5), 501-511. https://doi.org/10.1080/14703297.2021.1961097.

Latifi, S., Noroozi, O., & Talaee, E. (2021). Peer feedback or peer feedforward? Enhancing students' argumentative peer learning processes and outcomes. British Journal of Educational Technology, 52(2), 768-784. https://doi.org/10.1111/bjet.13054.

Latifi, S., Noroozi, O., & Talaee, E. (2020). Worked example or scripting? Fostering students' online argumentative peer feedback, essay writing and learning. Interactive Learning Environments, 1-15. https://doi.org/10.1080/10494820.2020.1799032.

Liljedahl, P. (2010): The Four Purposes of Assessment, *Vector* 2, 4-12

Mitchell, R. & Neill, M. (1992): *Criteria for evaluation of student assessment systems*, National Forum on Assessment

McTighe, J. & Ferrara, S. (1998): *Assessing learning in the classroom*, National Education Association

Noroozi, O. (2018). Considering students' epistemic beliefs to facilitate their argumentative discourse and attitudinal change with a digital dialogue game. Innovations in Education and Teaching International, 55(3), 357-365. https://doi.org/10.1080/14703297.2016.1208112.

Noroozi, O. (2022). The role of students' epistemic beliefs for their argumentation performance in higher education. Innovations in Education and Teaching International.1-12. https://doi.org/10.1080/14703297.2022.2092188.

Noroozi, O., Dehghanzadeh, H., & Talaee, E. (2020). A systematic review on the impacts of game-based learning on argumentation skills. Entertainment Computing, 35, 100369. https://doi.org/10.1016/j.entcom.2020.100369.

Noroozi, O., McAlister, S., & Mulder, M. (2016). Impacts of a digital dialogue game and epistemic beliefs on argumentative discourse and willingness to argue. *The International Review of Research in Open and Distributed Learning*, *17*(3). http://dx.doi.org/10.19173/irrodl.v17i3.2297.

Ortega-Sanchez, C. (2016): Written exams: How effectively are we using them? *Procedia – Social and Behavioral Sciences* 228, 144-148

Romagnano, L. (2001): Implementing the Assessment Standards: The Myth of Objectivity in Mathematics Assessment. *Mathematics Teacher* 94 (1), 31–37

Valero Haro, A., Noroozi, O., Biemans, H.J.A., & Mulder, M. (2019). First-and second-order scaffolding of argumentation competence and domain-specific knowledge acquisition: a systematic review. Technology, Pedagogy and Education, 28(3), 329-345. https://doi.org/10.1080/1475939X.2019.1612772.

Valero Haro, A, Noroozi, O., Biemans, H. J. A., & Mulder, M. (2022). Argumentation Competence: Students' argumentation knowledge, behavior and attitude and their relationships with domain-specific knowledge acquisition. Journal of Constructivist Psychology, 35(1),123-145. https://doi.org/ 10.1080/10720537.2020.1734995.

White, A. R. (2019): A Discussion and Students' Thoughts on the Assessment Methods used at the Tertiary Level. HAL-01976361